

Faster Approximate Pattern Matching in Compressed Repetitive Texts

Travis Gagie · Paweł Gawrychowski ·
Christopher Hoobin · Simon J. Puglisi

the date of receipt and acceptance should be inserted later

Abstract Motivated by the imminent growth of massive, highly redundant genomic databases, we study the problem of compressing a string database while simultaneously supporting fast random access, substring extraction and pattern matching to the underlying string(s). Bille et al. (2011) recently showed how, given a straight-line program with r rules for a string s of length n , we can build an $\mathcal{O}(r)$ -word data structure that allows us to extract any substring of length m in $\mathcal{O}(\log n + m)$ time. They also showed how, given a pattern p of length m and an edit distance $k \leq m$, their data structure supports finding all occ approximate matches to p in s in $\mathcal{O}(r(\min(mk, k^4 + m) + \log n) + \text{occ})$ time. Rytter (2003) and Charikar et al. (2005) showed that r is always at least the number z of phrases in the LZ77 parse of s , and gave algorithms for building straight-line programs with $\mathcal{O}(z \log n)$ rules. In this paper we give a simple $\mathcal{O}(z \log n)$ -word data structure that takes the same time for substring extraction but only $\mathcal{O}(z \min(mk, k^4 + m) + \text{occ})$ time for approximate pattern matching.

Keywords Compressed pattern matching · Approximate pattern matching · LZ77

1 Introduction

The recent revolution in high-throughput sequencing technology has made the acquisition of large genomic sequences drastically cheaper and faster. As the new technology takes hold, ambitious sequencing projects such as the 1,000 Human Genomes [4] and the 10,000 Vertebrate Genomes [10] projects are set to create large databases of strings (genomes) that vary only slightly from each other, and so will

A preliminary version of this work appeared in the proceedings of ISAAC 2011 [8].

T. Gagie
Department of Computer Science
Aalto University, Finland.
E-mail: travis.gagie@aalto.fi

P. Gawrychowski
Max Planck Institute
Saarbrücken, Germany
E-mail: gawry@cs.uni.wroc.pl

C. Hoobin
School of Computer Science and Information Technology
Royal Melbourne Institute of Technology, Australia
E-mail: christopher.hoobin@rmit.edu.au

S. J. Puglisi
Department of Informatics
King's College London, United Kingdom
E-mail: simon.puglisi@kcl.ac.uk

contain large numbers of long repetitions. Efficient storage of these collections is not enough: fast access to enable search and sequence alignment is paramount. The utility of such a data structure is not limited to the treatment of DNA collections. Ferragina and Manzini’s recent study of the compressibility of web pages reveals enormous redundancy in web crawls [5]. Exploiting this redundancy to reduce space while simultaneously enabling fast access and search over crawled pages (for snippet generation or cached page retrieval) is a significant challenge. The problem of compressing and indexing such highly repetitive strings (or string collections) was introduced in [20] (see also [15]). With an LZ78- or BWT-based data structure [1, 6] we can store a string s of length n in space bounded in terms of the t th-order empirical entropy [16], for any $t = o(\log_\sigma n)$, and later extract any substring of length m in $\mathcal{O}(m/\log_\sigma n)$ time. For very repetitive texts, however, compression based on the LZ77 [23] can use significantly fewer than $nH_t(s)$ bits [20].

Rytter [19] showed that the number z of phrases in the LZ77 parse of s is at most the number of rules in the smallest straight-line program (SLP) for s^1 . He then showed how the LZ77 parse can be turned into an SLP for s with $\mathcal{O}(z \log n)$ rules whose parse-tree has height $\mathcal{O}(\log n)$. This SLP can be viewed as a data structure that stores s in $\mathcal{O}(z \log n)$ words and supports substring extraction in $\mathcal{O}(\log n + m)$ time. Bille, Landau, Raman, Rao, Sadakane and Weimann [2] showed how, given an SLP for s with r rules, we can build a data structure that takes $\mathcal{O}(r)$ words and supports substring extraction in $\mathcal{O}(\log n + m)$ time regardless of the height of the parse tree. Unfortunately, since no polynomial-time algorithm is known to produce an SLP for s with $o(z \log n)$ rules, even with no bound on the height, we still do not know how, efficiently, to build a data structure that has better bounds than Rytter’s.

Bille et al. [2] also show how, given a pattern p of length m and an edit distance $k \leq m$, their data structure supports finding all occ approximate matches to p in s in $\mathcal{O}(r(\min(mk, k^4 + m) + \log n) + \text{occ})$ time. Their main idea is that, if there is a rule $X \rightarrow YZ$ in the SLP and we have already found all the approximate matches in expansions of Y and Z then, to find all the approximate matches in the expansion of X , we need only search the substring consisting of the $m+k$ last characters of Y ’s expansion concatenated with the first $m+k$ characters of Z ’s expansion. Extracting these characters with their data structure takes $\mathcal{O}(\log n + m)$ time per rule, or $\mathcal{O}(r(\log n + m))$ time in total. In this paper we discuss two improvements to this idea: first, by the same argument, we need only search the $m+k$ characters to either side of the phrase boundaries in the LZ77 parse; second, since we know in advance where those phrase boundaries are, we do not need the full power of random access. Our first observation immediately improves Bille et al.’s time bound for approximate matching to $\mathcal{O}(z(\min(mk, k^4 + m) + \log n) + \text{occ})$, while our second has led us to develop a data structure whose time bound is $\mathcal{O}(z \min(mk, k^4 + m) + \text{occ})$.

Neither Rytter’s nor Bille et al.’s data structures are practical. However, in another strand of recent work, Kreft and Navarro [12, 13] introduced a variant of LZ77 called LZ-End and gave a data structure based on it with which we can store s in $\mathcal{O}(z' \log n) + o(n)$ bits, where z' is the number of phrases in the LZ-End parse of s , and later extract any *phrase* (not arbitrary substring) in time proportional to its length. The $o(n)$ term can be removed at the cost of slowing extraction down by an $\mathcal{O}(\log n)$ factor. Extracting arbitrary substrings is fast in practice but could be slow in the worst case. Also, although the LZ-End encoding is small in practice for very repetitive strings, it is not clear whether z' can be bounded in terms of z .

Our Contribution. In this paper we describe a simple $\mathcal{O}(z \log n)$ -word data structure, which we call the *block graph* for s , that takes $\mathcal{O}(\log n + \ell - f)$ time to extract any substring $s[f..\ell]$ but lets us add bookmarks to speed up extraction from pre-specified points. This allows us to find all occ approximate matches of a pattern of length m in $\mathcal{O}(z \min(mk + m, k^4 + m) + \text{occ})$ time. Our space bound (in terms of z) and substring extraction time are the same as Bille et al.’s [2]; our approximate pattern matching time is faster both because we replace r by z (which, as noted above, they can too) and because we remove the $\log n$ term, which is due to the overhead for random access. More importantly, however, our results require much simpler machinery. We believe the block graph is the first practical data structure with solid theoretical guarantees for compression and retrieval of highly repetitive collections.

In the next section we describe the block graph. Then, in Section 3, we relate the size of the block graph to the size of the LZ77 parsing of its underlying string. We show that a block graph naturally compresses the string while allowing efficient random access and extraction of substrings. In Section 4 we show how to augment the block graph to support fast approximate pattern matching. In Section 6

¹ In this paper we consider only the version of LZ77 without self-referencing, sometimes called LZSS [21].

the substring $s[i..j]$. Notice that $s[i'..j']$ is completely contained within some block at depth d — this is one reason why we use overlapping blocks — and, since $s[i'..j']$ is the first occurrence of that substring in s , that block is associated with an internal node v . We replace the pointer from u to $\langle i..j \rangle$ by a pointer to v and the offset of i' in v 's block. For the example shown in Figure 1, $\langle 17..21 \rangle$ previously had children $\langle 17..20 \rangle$ and $\langle 19..21 \rangle$. The blocks $s[17..20] = abab$ and $s[19..21] = aba$, which first occur in positions 4 and 1, respectively. Therefore, we replace $\langle 17..21 \rangle$'s pointer to $\langle 17..20 \rangle$ by a pointer to $\langle 1..8 \rangle$ and the offset 3; we replace its pointer to $\langle 19..21 \rangle$ by another pointer to $\langle 1..8 \rangle$ and the offset 0.

Extracting a single character $s[i]$ in $\mathcal{O}(\log n)$ time is fairly straightforward: we start at the root and repeatedly descend to any child whose block contains $s[i]$; if we come to a leaf u such that $s[i]$ is the j th character in u 's block but, instead of pointing to a child whose block contains $s[i]$, u stores a pointer to internal node v and offset c , then we follow u 's pointer to v and extract the $(j + c)$ th character in v 's block; finally, when we arrive at an internal node with maximum depth, we report the appropriate character of its block, which is stored there explicitly. By definition the maximum depth of the block graph is $\log n$ and at each depth, we either descend immediately in $\mathcal{O}(1)$ time, or follow a pointer from a leaf to an internal node in $\mathcal{O}(1)$ time and then descend. Therefore, we use a total of $\mathcal{O}(\log n)$ time.

For example, suppose we want to extract the 11th character from $s = \text{abaababaabaababaababa}$ using the block graph shown in Figure 1. Starting at the root, we can descend to either child, since both their blocks contain $s[11]$; suppose we descend to the left child, $\langle 1..16 \rangle$. From $\langle 1..16 \rangle$ we can descend to either the middle or right children; suppose we descend to the right child, $\langle 9..16 \rangle$. Since $\langle 9..16 \rangle$ is a leaf, the pointer to child $\langle 9..12 \rangle$ has been replaced by a pointer to $\langle 1..8 \rangle$ and offset 0, while the pointer to child $\langle 11..14 \rangle$ has been replaced by another pointer to $\langle 1..8 \rangle$ and offset 2. This is because the first occurrence of $s[9..12] = \text{abaa}$ is $s[1..4]$ and the first occurrence of $s[11..14] = \text{aaba}$ is $s[3..6]$. Suppose we follow the second pointer. Since we would have extracted the first character from $\langle 11..14 \rangle$'s block, we are now to extract the third character from $\langle 1..8 \rangle$'s block. We can descend to either $\langle 1..4 \rangle$ and extract the third character of its block, or descend to $\langle 3..6 \rangle$ and extract the first character of its block.

Extracting longer substrings is similar, but complicated by the fact that we want to avoid breaking the substring into too many pieces as we descend. In the next section we will show how to extract any substring of length m in $\mathcal{O}(\log n + m)$ time; however, we first prove an upper bound on the block graph's size.

3 Fast random access in compressed space

In this section we show that block graphs achieve compression while simultaneously allowing easy access to the underlying string. Our space result relies on the following easily proved lemma.

Lemma 1 ([7]) *The first occurrence of any substring in s must touch at least one boundary between phrases in the LZ77 parse.*

Lemma 1 allows us to relate the size of the block graph to the LZ77 parsing of the underlying string, as summarized below.

Theorem 1 *The block graph for s takes $\mathcal{O}(z \log^2 n)$ bits.*

Proof Each internal node's block is the first occurrence of that substring in s so, by Proposition 1, it must touch at least one boundary between phrases in the LZ77 parse. Since each such boundary can touch at most three blocks in the same level, there are at most $3z$ internal nodes in each level. It follows that there are $\mathcal{O}(z \log n)$ nodes in all. Since each node stores $\mathcal{O}(\log n)$ bits, the whole block graph takes $\mathcal{O}(z \log^2 n)$ bits.

We define the query $\text{extract}(u, i, j)$ to return the i th through j th characters in u 's block. Notice that, if u is the root, then these characters are $s[i..j]$. We now show how to implement extract queries in such a way that extracting a substring of s with length m takes $\mathcal{O}(\log n + m)$ time.

There are three cases to consider when performing $\text{extract}(u, i, j)$: u could be an internal node at maximum depth, in which case we simply return the i th through j th characters of its block, which are stored explicitly; u could be an internal node with children; or u could be a leaf. First suppose that u is an internal node with children. Let d be u 's depth and $b = 2^{\lceil \log_2 n \rceil - d}$; notice b is the length of u 's block

unless the block is a suffix of s , in which case the block might be shorter. If the interval $[i..j]$ is completely contained in one of the intervals $[1..b/2]$, $[b/4 + 1..3b/4]$ or $[b/2 + 1..b]$, then we set v to be the left, middle or right child of u , respectively (choosing arbitrarily if two intervals each completely contain $[i..j]$), and implement $\text{extract}(u, i, j)$ as either $\text{extract}(v, i, j)$, $\text{extract}(v, i - b/4, j - b/4)$ or $\text{extract}(v, i - b/2, j - b/2)$. Otherwise, $[i..j]$ must be more than a quarter of $[1..b]$ and we can split $[i..j]$ into 2 or 3 subintervals, each of length at least $b/8$ but completely contained in one of $[1..b/2]$, $[b/4 + 1..3b/4]$ or $[b/2 + 1..b]$; this is the other reason why we use overlapping blocks. We implement $\text{extract}(u, i, j)$ with an extract query for each subinterval.

Now suppose that u is a leaf. Again, let d be u 's depth and $b = 2^{\lceil \log_2 n \rceil - d}$. If the interval $[i..j]$ is completely contained in one of the intervals $[1..b/2]$, $[b/4 + 1..3b/4]$ or $[b/2 + 1..b]$, then we set v to be the first, second or third internal node at the same depth to which u points, respectively, and implement $\text{extract}(u, i, j)$ as $\text{extract}(v, i', j')$, where i' and j' are i and j plus the appropriate offset. Otherwise, $[i..j]$ must be more than a quarter of $[1..b]$; we split $[i..j]$ into subintervals and implement $\text{extract}(u, i, j)$ with an extract query for each subinterval, as before.

Theorem 2 *Extracting a substring $s[f..\ell]$ from the block graph of s takes $\mathcal{O}(\log n + \ell - f)$ time.*

Proof Consider the query $\text{extract}(\text{root}, f, \ell)$ and let d be the first depth at which we split the interval. Descending to depth d takes a total of $\mathcal{O}(d)$ time. By induction, if we perform a query $\text{extract}(v, i, j)$ on a node v at depth $d' > d$, then $j - i + 1$ is more than a quarter of the block size $2^{\lceil \log_2 n \rceil - d'}$ at that level. It follows that we make $\mathcal{O}((\ell - f + 1)/2^{\log n - d'})$ calls to extract at depth d' , each of which takes $\mathcal{O}(1)$ time. Summing over the depths, we use a total of $\mathcal{O}(\log n + \ell - f)$ time. \square

One interesting property of our block graph structure is that, at the cost of storing a node for every possible block of size $n/2^d$ — i.e., storing $\mathcal{O}(2^d \log n)$ extra bits — we can remove the top d levels and, thus, change the overall space bound to $\mathcal{O}(z(\log n - d) \log n + 2^d \log n)$ bits and reduce the access time to $\mathcal{O}(\log n - d)$. For example, if $d = \log z$, then we store a total of $\mathcal{O}(z \log n \log(n/z))$ bits and need only $\mathcal{O}(\log(n/z))$ time for access. If $d = \log(n/\log^2 n)$, then we store a total of $\mathcal{O}(z \log n \log \log n + n/\log n)$ bits and reduce the access time to $\mathcal{O}(\log \log n)$.

González and Navarro [11] showed how, by applying grammar-based compression to a difference-coded suffix array (SA), we can build a new kind of compressed suffix array that supports access to $\text{SA}[i..j]$ in $\mathcal{O}(\log n + \ell - f)$ time. It seems likely that, by using a modified block graph of the difference-coded suffix array instead of a grammar, we can improve their access time to $\mathcal{O}(\log \log n + \ell - f)$ at the cost of only slightly increasing their space bound.

4 Accelerated approximate pattern matching

Suppose we are given an uncompressed string s of length n , the LZ77 parse [23] of s , a pattern p of length $m \leq n$ and an edit distance $k \leq m$. The primary matches of p are the substrings of s within edit distance k of p whose characters are all within distance $(m + k)$ of phrase boundaries in the parse. It is not difficult to find all p 's primary matches in $\mathcal{O}(z \min(mk + m, k^4 + m))$ time, where z is the number of phrases. To do this, we extract the substrings all of whose characters are within distance $(m + k)$ of phrase boundaries and apply to them either the sequential approximate pattern-matching algorithm by Landau and Vishkin [14] or the one by Cole and Hariharan [3].

Once we have found p 's primary matches, we can use them to find the approximate matches not within distance $(m + k)$ of any phrase boundary, which are called p 's secondary matches. To do this, we process the phrases from left to right, maintaining a sorted list of the approximate matches we have already found. For each phrase copied from a previous substring $s[i..j]$, we search in the list to see if there are any approximate matches in $s[i..j]$ that are not completely contained in $s[i..i + m + k - 1]$ or $s[j - m - k + 1..j]$. If there are, we insert the corresponding secondary matches in our list. Processing all the phrases takes $\mathcal{O}(z + \text{occ})$ time, where occ is the number of approximate matches to p in s . Notice that finding p 's secondary matches does not require access to s .

As noted in Section 1, Bille et al. [2] showed how, given a straight-line program for s with r rules, we can build an $\mathcal{O}(r)$ -word data structure that allows us to extract any substring $s[f..\ell]$ in $\mathcal{O}(\log n + \ell - f)$ time. When the straight-line program is built with the best known algorithm for approximately minimizing the number of rules, $r = \mathcal{O}(z \log n)$ [19]. It follows that we can store s in $\mathcal{O}(z \log n)$ words such that,

given p and k , in $\mathcal{O}(z(\log n + m))$ time we can extract all the characters within distance $(m + k)$ of phrase boundaries and, therefore, find all p 's approximate matches in $\mathcal{O}(z(\min(mk + m, k^4 + m) + \log n) + \text{occ})$ time. (Bille et al. themselves gave a bound of $\mathcal{O}(r(\min(mk + m, k^4 + m) + \log n) + \text{occ})$ but, since even the smallest straight-line program for s has at least z rules [19], the one we state is slightly stronger.)

The key to supporting approximate pattern matching in the block graph is the addition of *bookmarks*, which will allow us to quickly extract certain regions of the underlying string. To add a bookmark to a character $s[i]$, for each block size b in the block graph, we store pointers to the two nodes whose blocks of size $2b$ completely contain the first occurrence of the substrings $s[i - b + 1..i]$ and $s[i..i + b - 1]$, and those occurrences' offsets in the blocks. Thus, storing a bookmark takes $\mathcal{O}(\log n)$ words. To extract a substring that touches $s[i]$, we extract, separately, the parts of the substring to the left and right of $s[i]$. Without loss of generality, we assume the part $s[i..j]$ to the right is longer and consider only how to extract it. We first find the smallest block size $b \geq j - i + 1$, then follow the pointer to the node whose block of size $2b$ contains the first occurrence $s[i..i + b - 1]$. Since that node has height $\mathcal{O}(\log(j - i + 1))$, we can extract $s[i..j]$ in $\mathcal{O}(j - i + 1)$ time.

Lemma 2 *Extracting a substring $s[f..\ell]$ that touches a bookmark takes $\mathcal{O}(\ell - f)$ time.*

Inserting a bookmark to each phrase boundary in the LZ77 parse takes $\mathcal{O}(z \log n)$ words and allows us, given m and k , to extract the characters within distance $(m + k)$ of phrase boundaries in a total of $\mathcal{O}(zm)$ time. Combined with the approach described above for finding secondary occurrences, we have our main result.

Theorem 3 *Let s be a string of length n whose LZ77 parse consists of z phrases. We can store s in $\mathcal{O}(z \log n)$ words such that, given a pattern p of length $m \leq n$ and an edit distance $k \leq m$, we can find all occ substrings of s within edit distance k of p in $\mathcal{O}(z \min(mk + m, k^4 + m) + \text{occ})$ time.*

Note that, in the above theorem, the time to find all p 's approximate matches is the same as if we were keeping s uncompressed, as in the approach described at the start of this section.

We note in passing that we can combine our results with those of Kreft and Navarro [13] to obtain a new worst-case upper bound for LZ77-based indexing. Specifically, replacing their data structures for access to the string by a block graph with a bookmark at each phrase boundary, and replacing two of their other data structures by faster (and larger, but still $\mathcal{O}(z \log^2 n)$ bits) data structures, we can store s in $\mathcal{O}(z \log^2 n)$ bits such that, given a pattern p of length m , we can find all occurrences of p in s in $\mathcal{O}(m^2 + (m + \text{occ}) \log \log z)$ time. Their index is practical but potentially larger and slower in the worst case.

5 Efficient representation of block graphs

We now describe an implementation of block graphs which is efficient in practice. The main idea is to represent the shape of the graph (the internal nodes and their pointers) using bitvectors and operations from succinct data structures, and to carefully allocate space for the leaf nodes depending on their distance from the root. Below we make use of two familiar operations for bitvectors: *rank* and *select*. Given a bitvector B , a position i , and a type of bit b (either 0 or 1), $\text{rank}_b(B, i)$ returns the number of occurrences of b before position i in B and $\text{select}_b(B, i)$ returns the position of the i th b in B . Efficient data structures supporting these operations have been extensively studied (see, e.g. [17, 18]).

Each level of the block graph consists of a number of nodes, either internal nodes, or leaves. Let B_d be a bitvector which says whether the i th node (from the left) at depth d is a leaf, $B_d[i] = 0$, or an internal node $B_d[i] = 1$. We define another bitvector R_d , where $R_d[i] = 1$ if and only if $B_d[i] = 1$ and $B_d[i + 1] = 1$ for $i < n - 1$. That is, we mark a 1 bit for each instance of two adjacent internal nodes in B_d , otherwise $R_d[i] = 0$. Let L_d be an array that holds leaf nodes at depth d . The structure of a leaf node is discussed below. Finally, let T be the concatenation of the textual representation (ie. the corresponding substrings) of all internal nodes at the truncated depth.

Navigating the block graph. The main operation is to traverse from an internal node to one of its three children. Say we are currently at the j th internal node at depth d of the block graph — that is, we are at $B_d[i]$, where $i = \text{select}_1(B_d, j)$. Each internal node has three children. If these children were independent

then locating the left child of the current node would be simply three times the node’s position on its level, that is $3j = 3 \cdot \text{rank}_1(B_d, i)$. However, in a block graph adjacent internal nodes share exactly one child, so we correct for this by subtracting the number of adjacent internal nodes at this depth prior to the current node — this is given by $\text{rank}_1(R_d, i)$. To find the position corresponding to the left child of a node in B_{d+1} we compute

$$\text{leftchild}(B_d, i) = 3\text{rank}_1(B_d, i) - \text{rank}_1(R_d, i)$$

Given the address of the left child it is easy to find the center or right child by adding 1 or 2 respectively to the result of `leftchild`. If $B_d[i] = 0$ then we are at a leaf node. Intuitively, to access its leaf information in L_d we call $L_d[\text{rank}_0(B_d, i)]$. Once we reach the truncated depth to access the text of an internal node we compute its offset in T , $T[(\text{rank}_1(B_d, i) * \text{truncated length})]$.

Leaf nodes. In a block graph leaves point to internal nodes. For each leaf we store two values, the position of the destination node on the current level, and an offset in the destination node pointing to the beginning of the leaf block. Note that we do not need to store the depth of the destination node. It is, by definition, on the level above the leaf, and we know this by keeping keep track of the depth during each step in a traversal. To improve compression we store leaf positions and offsets in two separate arrays. At depth d there are no more than $2^{d+1} - 1$ possible nodes, so we can store each position in $\log(2^{d+1} - 1)$ bits. Given that the length of a node at depth d is $b = 2^{\lceil \log n \rceil - d}$ and leaf nodes point to an internal node on the level above, we store each offset in $\log(2^{\lceil \log n \rceil - d - 1})$ bits.

6 Experiments

We have developed an implementation of block graphs² and tested it on the real-world texts of the Pizza-Chili Repetitive Corpus³, a standard testbed for data structures designed for repetitive strings.

We compared compression achieved by the block graph to the LZ-End data structure by Kreft and Navarro [12], and to the general-purpose compressors `gzip` and `7zip`; the results are shown in Table 1. We used `gzip` and `7zip` with the settings `-9` and `-t7z -m0=lzma -mx=9 -mfb=64 -md=32m -ms=on`, respectively, while LZ-End was executed with its default settings. Throughout our experiments all block graphs were truncated such that the smallest blocks each took 4 bytes. Note that `gzip` and `7zip` provide compression only, not random access, and are included as reference points for achievable compression.

We then compared how quickly block graphs and LZ-End support extracting substrings of various lengths; the results are shown in Figure 2. Each run of extractions was performed across 10,000 randomly-generated queries. Experiments were conducted on an Intel Core i7-2600 3.4 GHz processor with 8GB of main memory, running Linux 3.3.4; code was compiled with GCC version 4.7.0 targeting x86_64 with full optimizations. Caches were dropped between runs with `sync && echo 1 > /proc/sys/vm/drop_caches`.

Although `7zip` achieves much better compression block graphs achieve better compression than `gzip` except on the *Escherichia Coli* and *influenza* files. Most importantly, our experiments show that block graphs generally achieve compression comparable to that achieved by LZ-End while supporting significantly faster substring extraction.

7 Conclusions

Efficient storage and retrieval of highly repetitive strings, and approximate pattern matching in them, are important tools in bioinformatics and will become even more important as genomic databases grow. In this paper we have presented a new data structure, the *block graph*, that stores highly repetitive strings in compressed space, supports random access in reasonable time and supports extraction from pre-specified points much faster. Our analysis and experiments show that the block graph is competitive both in theory and in practice.

² Available at <http://www.github.com/choobin/block-graph>

³ <http://pizzachili.dcc.uchile.cl/rep corpus.html>

Table 1 Size in bytes of repetitive corpus files encoded with ASCII, gzip, 7zip, LZ-End and block graphs.

Collection	ASCII	gzip	7zip	LZ-End	Block graph
Escherichia Coli	112,689,515	31,535,023	6,147,962	49,106,638	49,716,456
cere	461,286,644	120,834,282	6,077,972	41,342,784	57,689,376
coreutils	205,281,778	49,920,838	3,999,812	35,863,520	47,795,692
einstein.en.txt	467,626,544	163,664,285	323,779	2,247,204	3,969,392
influenza	154,808,555	10,636,899	2,111,974	21,507,089	33,171,036
kernel	257,961,616	69,396,104	2,087,006	19,347,734	24,045,332
para	429,265,758	116,073,220	8,117,573	57,415,176	72,393,196
world leaders	46,968,181	8,287,665	606,438	4,525,317	7,321,720

Acknowledgments

Many thanks to Francisco Claude, Juha Kärkkäinen, Sebastian Kreft, Gonzalo Navarro, Jorma Tarhio and Alexandru Tomescu, for helpful discussions.

References

1. Arroyuelo D, Navarro G, Sadakane K (2012) Stronger Lempel-Ziv based compressed text indexing. *Algorithmica* 62(1–2)
2. Bille P, Landau GM, Raman R, Sadakane K, Satti SR, Weimann O (2011) Random access to grammar-compressed strings. In: *Proceedings of the 22nd Symposium on Discrete Algorithms (SODA)*, pp 373–389
3. Cole R, Hariharan R (2002) Approximate string matching: A simpler faster algorithm. *SIAM Journal on Computing* 31(6):1761–1782
4. Durbin R, et al (2010) 1000 genomes. <http://www.1000genomes.org/>
5. Ferragina P, Manzini G (2010) On compressing the textual web. In: *Proceedings of the 3rd Conference on Web Search and Data Mining (WSDM)*, pp 391–400
6. Ferragina P, Venturini R (2007) A simple storage scheme for strings achieving entropy bounds. *Theoretical Computer Science* 372(1):115–121
7. Gagie T, Gawrychowski P (2010) Grammar-based compression in a streaming model. In: *Proceedings of the 4th Conference on Language and Automata Theory and Applications (LATA)*, pp 273–284
8. Gagie T, Gawrychowski P, Puglisi SJ (2011) Faster approximate pattern matching in compressed repetitive texts. In: *Proceedings of the 22nd International Symposium on Algorithms and Computation (ISAAC)*, pp 653–662
9. Gagie T, Gawrychowski P, Kärkkäinen J, Nekrich Y, Puglisi SJ (2012) A faster grammar-based self-index. In: *Proceedings of the 6th Conference on Language and Automata Theory and Applications (LATA)*, pp 240–251
10. Genome 10K Community of Scientists (2009) A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *Journal of Heredity* 100:659–674
11. González R, Navarro G (2007) Compressed text indexes with fast locate. In: *Proceedings of the 18th Symposium on Combinatorial Pattern Matching (CPM)*, pp 216–227
12. Kreft S, Navarro G (2010) LZ77-like compression with fast random access. In: *Proceedings of the Data Compression Conference (DCC)*, pp 239–248
13. Kreft S, Navarro G (2011) Self-indexing based on LZ77. In: *Proceedings of the 22nd Annual Symposium on Combinatorial Pattern Matching (CPM)*, pp 41–54
14. Landau GM, Vishkin U (1989) Fast parallel and serial approximate string matching. *Journal of Algorithms* 10(2):157–169
15. Mäkinen V, Navarro G, Sirén J, Valimäki N (2010) Storage and retrieval of highly repetitive sequence collections. *Journal of Computational Biology* 17(3):281–308
16. Manzini G (2001) An analysis of the Burrows-Wheeler transform. *Journal of the ACM* 48(3):407–430
17. Okanohara D, Sadakane K (2007) Practical entropy-compressed rank/select dictionary. In: *Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX)*
18. Raman R, Raman V, Satti SR (2007) Succinct indexable dictionaries with applications to encoding k -ary trees, prefix sums and multisets. *ACM Transactions on Algorithms* 3(4)

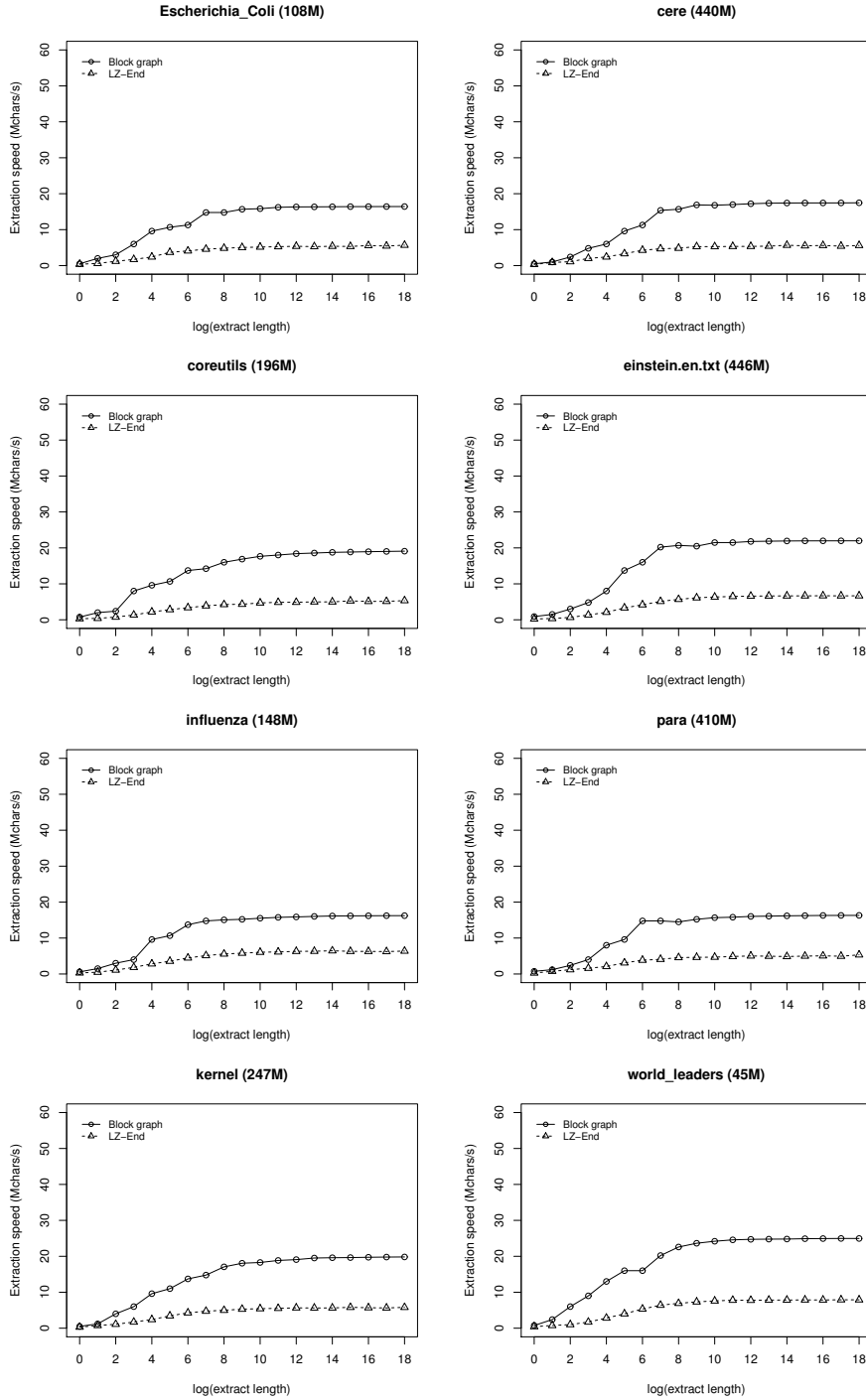


Fig. 2 Random access and extraction speeds. Times are averaged over 10,000 random substring extractions.

19. Rytter W (2003) Application of Lempel-Ziv factorization to the approximation of grammar-based compression. Theoretical Computer Science 302(1-3):211–222
20. Sirén J, Välimäki N, Mäkinen V, Navarro G (2008) Run-length compressed indexes are superior for highly repetitive sequence collections. In: Proceedings of the 15th Symposium on String Processing and Information Retrieval (SPIRE), pp 164–175
21. Storer JA, Szymanski TG (1982) Data compression via textual substitution. Journal of the ACM 29(4):928–951

22. Vezzi F, Del Fabbro C, Tomescu AI, Policriti A (2012) rNA: a fast and accurate short reads numerical aligner. *Bioinformatics* 28(1):123–124
23. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* 23(3):337–343